

Advantages of the Extended Water Flow Algorithm for Handwritten Text Segmentation

Darko Brodić

University of Belgrade, Technical Faculty Bor,
V. J. 12, 19210 Bor, Serbia
dbrodic@tf.bor.ac.rs

Abstract. This paper identifies the advantages of the specific approach to water flow algorithm for multi-skewed handwritten text line segmentation. Original water flow algorithm assumes that hypothetical water flows, from both left and right sides of the document image frame, face obstruction from part of character, character, and group of characters in text lines. The stripes of areas left unwetted on the document image frame are finally labeled for the extraction of text lines. However, the method defines parameter water flow angle for flooding which depends on the text line slopes of each specific document. The estimation of the appropriate parameter value is difficult and limited as well. The limitation is manifested by possible election of only 4 values for this parameter. Extended approach has introduced enlargement of the parameter range. Consequently, decision making and the selection of the small values of the parameter below the minimum given by the original method shows improvement in the handwritten text line segmentation process. It is confirmed by the measurement on different types of letters.

Keywords: Image processing; Document image processing; Text line segmentation; Handwritten text; Water flow algorithm.

1 Introduction

Text line segmentation of handwritten text is a key element of the optical character recognition [1]. It implies a labeling process that assign the same label to spatially aligned units, connected components or characteristic points [1]. Further, based on the obtained labels, text is divided into different regions each one representing text line. After text line segmentation is finished, it provides the essential information for the consecutive documents image steps such as baseline detection, skew identification and correction, other text feature extraction and character recognition.

Related work on text line segmentation can be categorized in few directions [1,2]: projection based methods, Hough transform methods, grouping methods, methods for processing overlapping and touching components, stochastic methods, smearing methods, and others method.

In smearing methods the consecutive black pixels along the horizontal direction are smeared [3]. Consequently, the white space between black pixels is filled

with black pixels. It is valid only if their distance is within a predefined threshold. This way, enlarged area of black pixels around text is formed. It is so-called boundary growing area. Hence, the smeared image encloses text lines.

As mentioned many methods can be employed for text line segmentation, but smearing algorithms have the benefits. They are efficient and computationally inexpensive. Algorithm proposed in [4] is classified as smearing method. It assumed a hypothetical flow of water in a particular direction across image frame in a way that it faces obstruction from the characters of the text lines. As a result of algorithm, unwetted image frames are extracted. These areas are of major importance for text line segmentation. In our paper, this algorithm is extended as well as in [5,6] and further investigated, improved and adapted.

The rest of the paper is organized as follows: Section 2 describes basic water flow algorithm and introduces extended approach to water flow algorithm. Section 3 defines testing process and shows measurement results including comparative analysis. Section 4 makes conclusions.

2 Water Flow Algorithm

Document text image is obtained as result of scanning process. It is grayscale text image described by intensity matrix $D(i,j) \in \{0, \dots, 255\}$, where $i \in \{1, M\}$ and $j \in \{1, N\}$ [7]. After applying intensity segmentation with binarization, it is converted into binary image $X(i,j)$. Hence, where $D(i,j) \geq D_{th}$, then $X(i,j)$ is 1, elsewhere $X(i,j)$ is 0. Consequently, D_{th} represents the threshold sensitivity decision value. It is obtained by global [8] or local binarization method [9,10,11].

Basic water flow algorithm [4] assumes hypothetical water flows under few angles of the document image frame from left to right and vice versa. In this hypothetically assumed situation, water is flowing across the image frame. Areas that are not wetted form unwetted ones. The stripes of unwetted areas are labeled for the extraction of text lines. Further, this hypothetical water flow is expected to fill up the gaps between text elements in the same text lines. Hence, unwetted areas left on the image frame lies under the text lines. Once the labeling is completed, the image is divided into two different types of stripes. First one contains text lines. The other one contains space between text lines. Furthermore, water flow angle α , as shown in Fig.1, is introduced as a referent parameter in [4]. This angle is measured between two lines which intersect each other at the endpoint of an obstacle. However, it is formed by labeling original document image using spatial filter mask. These masks for the water flows from left to right are shown in Fig.1.

Accordingly, algorithm creates unwetted areas under fixed water flow angles from the set $\{14^\circ, 18^\circ, 26.6^\circ, 45^\circ\}$ [4]. Using of the spatial filter masks is defined by the position of the text black pixels. They represent prospective seed points. If the pixel represents corner one, then the spatial filter mask will be exploited. Thus, it activates the algorithm process. As the result, unwetted stripes bounded text. This circumstance for $\alpha = 26.6^\circ$ is shown in Fig.2.

In the extended approach, originally extraction of the rectangular bounding box over the text objects is prerequisite. Further, they represent control areas

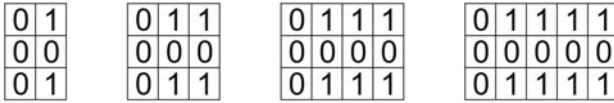


Fig. 1. Spatial filter mask defined by water flow angle: $\alpha = 45^\circ$ (left), $\alpha = 26.6^\circ$ (middle left), $\alpha = 18^\circ$ (middle right), and $\alpha = 14^\circ$ (right)

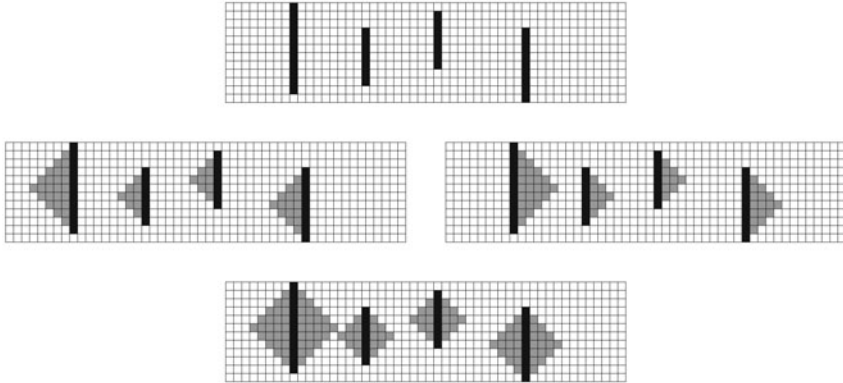


Fig. 2. Water flow algorithm with water flow angle $\alpha = 26.6^\circ$: Initial sample text containing four I letters (top), unwetted areas made by water flow from left to right (middle left), unwetted areas made by water flow from right to left (middle right), and united unwetted areas (bottom).

authorized for the extended water flow algorithm criteria activation. Bounding box is a rectangular region whose edges are parallel to the coordinate axes. It is defined by its endpoints: x_{min} , y_{min} , x_{max} , and y_{max} . Hence, each pixel $X(i,j)$ that belongs to the bounding box is given by [12]:

$$X(i,j) \mid (x_{min} \leq i \leq x_{max}) \wedge (y_{min} \leq j \leq y_{max}). \tag{1}$$

Inclusion of the point $X(i,j)$ in the bounding box is tested by verifying these four inequalities. If any one of them fails, then the point is not inside [13]. All text objects like letters, part of words or words are surrounded by bounding boxes [14]. It should be noted that from this point all actions are referred only to text in each independent bounding box.

Apart from original algorithm procedure, unwetted areas could be determined by lines. Each line is defined as:

$$y = ax + b, \tag{2}$$

where slope $a = \tan(\alpha)$. Two lines defined by angle α make connection in specific pixel creating closed unwetted area [5,6].

Modification made on water flow algorithm is in its definition. Consequently, it forms the water flow function which determines the water flow angle α . Still, making straight lines from boundary pixel type and connecting each others in specified point makes unwetted region as well. Currently, modified water flow algorithm is open to choose different water flow angle from 0° to 90° . In addition, different decision-making process on boundary text pixels is made. Fig.3 illustrates these circumstances [5,6].

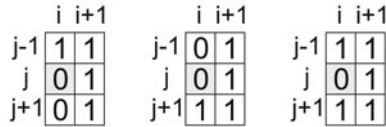


Fig. 3. Pixel type determination: upper boundary pixel (left), lower boundary pixel (middle), and boundary pixel for additional investigation (right)

Due to pixel type, the slope is $-\alpha$ or $+\alpha$. However, pixel without complete location has been additionally investigated. It depends on neighbor area of pixels. Apart from [5], enlarged window composed of $R \times S$ pixels, defined as a basis. In [6] $R = 5$ and $S = 7$ is proposed for complete investigation.

The main achievement of unwetted area is text line segmentation. However, the problem originates in broken words of text lines. Algorithm should contribute to join those words by unwetted areas. Those areas can be lengthen by using smaller angle α . This is represented in Fig.4.

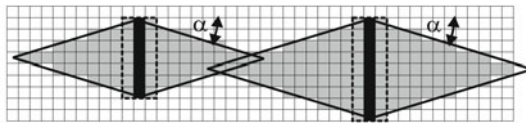


Fig. 4. Extended water flow algorithm involving water flow function $y = ax + b$ defining and bounding boxes surrounding text (box with dotted lines)

3 Testing, Results and Comparative Analysis

The proposed approach was tested using a group of binary handwritten documents images. These documents were written in English as well as in Serbian Latin and Cyrillic letters. The total number of analyzed handwritten text lines was 220. For the sake of conformity, documents body is the only considered in the analysis of the line segmentation.

Table 1 presents the number of correctly segmented text lines as well as different types of segmented errors. Errors are divided in 3 groups: split lines error, joined lines error, and lines including outlier words [15].

Split lines error represents the text lines which are wrongly divided by algorithm in two or more components [15]. Joined lines error corresponds to the situation where the sequence of n consecutive lines is considered by the algorithm as a unique line. In this situation $n - 1$ line represent the group of the erroneous [15]. Lines including outlier words correspond to lines containing words that are incorrectly assigned to two adjacent lines [15].

Table 1. Text lines segmentation results produced by original method with $\alpha = 14^\circ$ [4] and extended method with from the set $\{8^\circ, 10^\circ, 12^\circ\}$.

Table 1. Text lines segmentation results produced by original method with $\alpha = 14^\circ$ [4] and extended method with α from the set $\{8^\circ, 10^\circ, 12^\circ\}$

Water flow angle	14°		12°		10°		8°	
	#	%	#	%	#	%	#	%
Correctly segmented	88	40.00%	96	43.64%	144	65.45%	180	81.82%
Split lines error	132	60.00%	124	56.36%	76	34.55%	40	18.18%
Joined line error	0	0%	0	0%	0	0%	0	0%
Lines including outlier words	0	0%	0	0%	0	0%	0	0%
Sum	220	100%	220	100%	220	100%	220	100%

One of the representative results from test samples is given in Fig. 5.

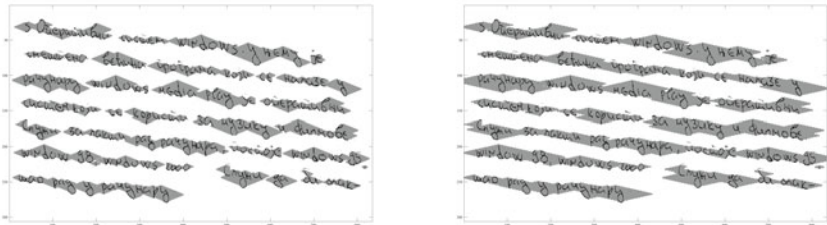


Fig. 5. Application of the algorithm to the handwritten sample text: $\alpha = 14^\circ$ (left), and $\alpha = 8^\circ$ (right)

From the above results the use of smaller water flow angle contributes to the better text segmentation results. Especially, water flow angles below 10° give segmentation results improvement. Hence, obtained results confirmed advantage of the extended over basic approach for the water flow algorithm.

4 Conclusions

In this paper, an extended approach to water flow algorithm for text line segmentation is presented. Water flow algorithm assumes the flow of the hypothetical water under few specified angles to the image frame from left to right and vice

versa. Instead of using limited choices of parameter called water flow angle, new approach introduced approach to freely choose this parameter. Furthermore, choosing smaller water flow angles contributes to the enlarged unwetted area which leads to better text line segmentation results. It is upheld by testing with the group of handwritten documents in Serbian and English language. Hence, the extended approach of the water flow algorithm confirms better text line segmentation ability.

References

1. Likforman-Sulem, L., Zahour, A., Taconet, B.: Text Line Segmentation of Historical Documents: A Survey. *IJDAR* 9(2-4), 123–138 (2007)
2. Amin, A., Wu, S.: Robust Skew Detection in Mixed Text/Graphics Documents. In: 8th ICDAR, Seoul, Korea, pp. 247–251 (2005)
3. Shi, Z., Govindaraju, V.: Line Separation for Complex Document Images using Fuzzy Runlength. In: *Int. Workshop on Document Image Analysis for Libraries*, Palo Alto, USA, pp. 306–312 (2004)
4. Basu, S., et al.: Text Line Extraction from Multi-skewed Handwritten Documents. *Pattern Recognition* 40(6), 1825–1839 (2007)
5. Brodic, D., Milivojevic, Z.: An Approach to Modification of Water Flow Algorithm for Segmentation and Text Parameters Extraction. In: *Camarinha-Matos, L.M., Pereira, P., Ribeiro, L. (eds.) IFIP AICT*, vol. 314, pp. 323–331. Springer, Heidelberg (2010)
6. Brodic, D., Milivojevic, Z.: Text Line Segmentation by Adapted Water Flow Algorithm. In: 10th Symposium NEUREL, Belgrade, Serbia, pp. 225–229 (2010)
7. Gonzalez, R.C., Woods, R.E.: *Digital Image Processing*, 2nd edn. Prentice-Hall, Englewood Cliffs (2002)
8. Otsu, N.: A Threshold Selection Method from Gray-level Histograms. *IEEE Tran. on Systems, Man, and Cybernetics* 9(1), 62–66 (1979)
9. Sauvola, L., Pietikainen, M.: Adaptive Document Image Binarization. *Pattern Recognition* 33(2), 225–236 (2000)
10. Bukhari, S.S., Shafait, F., Bruesl, T.M.: Adaptive Binarization of Unconstrained Hand-held Camera-captured Document Images. *J. of Universal Computer Science* 15(18), 3343–3363 (2009)
11. Khashman, A., Sekeroglu, B.: Document Image Binarisation Using a Supervised Neural Network. *Int. J. of Neural Systems* 18(5), 405–418 (2008)
12. Preparata, F.P., Shamos, M.I.: *Computational Geometry: An Introduction*. Springer, Berlin (1985)
13. Chang, C.M.: Detecting Ellipses via Bounding Boxes. *Asian Journal of Health and Information Sciences* 1(1), 73–84 (2006)
14. Wang, J., Leung, M.K.H., Hui, S.C.: Cursive Word Reference Line Detection. *Pattern Recognition* 30(3), 503–511 (1997)
15. Sanchez, A., et al.: Text Line Segmentation in Images of Handwritten Historical Documents. In: 1st IPTA, Sousse, Brasil, pp. 1–6 (2008)